DOCUMENT RESUME

ED 274 707 TM 860 584

AUTHOR Fuchs, Douglas
TITLE You Can Take a Test Out of the Situation, but You Can't Always Take the Situation Out of a Test: Bias in Minority Assessment.
PUB DATE [Jun 85]
NOTE 23p.; Portions of this paper were presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 1986) and at the Biennial Conference on Minority Assessment (2d, Tucson, AZ, November 1985).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; Blacks; Effect Size; Elementary Secondary Education; *Ethnic Groups; *Examiners; Hispanic Americans; Individual Testing; *Meta Analysis; *Minority Groups; Racial Factors; *Test Bias; *Testing Problems; Whites

ABSTRACT

This article presents a quantitative synthesis of examiner familiarity effects on Caucasian and minority students' test performance. Fourteen controlled studies were coded in terms of methodological quality (high vs. low) and race-ethnicity (Caucasian vs. Black and Hispanic). An analogue to analysis of variance conducted on weighted unbiased effect sizes indicated examiner familiarity produced a significant effect, with Caucasian and minority examinees' test performance raised by .05 and .72 standard deviations, respectively. Examiner familiarities differential effect on Caucasian and minority examinees did not interact with the methodological quality of the studies. Implications for test practice and research are discussed. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED274707

You Can Take A Test Out Of The Situation, But You Can't Always Take The Situation Out Of A Test: Bias In Minority Assessment

Douglas Fuchs

Peabody College of Vanderbilt University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. Fuchs

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Portions of this paper were presented at the annual meeting of the American Psychological Association, Washington, DC, August 1986 and at the Second Biennial Conference on Minority Assessment, Tucson, AZ, November 1985.

Requests for reprints should be sent to Douglas Fuchs, Box 328, Department of Special Education, Peabody College, Vanderbilt University, Nashville, TN 37203.

Running head: Minority Assessment

TM 860 584

BEST COPY AVAILABLE

Abstract

This article presents a quantitative synthesis of examiner familiarity effects on Caucasion and minority students' test performance. Fourteen controlled studies were coded in terms of methodological quality (high vs. low) and race-ethnicity (Caucasion vs. black and Hispanic). An analogue to analysis of variance conducted on weighted unbiased effect sizes (UES) indicated examiner familiarity produced a significant effect, with Caucasion and minority examinees' test performance raised by .05 and .72 standard deviations, respectively. Examiner familiarity's differential effect on Caucasion and minority examinees did not interact with the methodological quality of the studies. Implications for test practice and research are discussed.

You Can Take A Test Out Of The Situation, But You Can't Always Take The Situation Out Of A Test: Bias In Minority Assessment

Nearly two decades ago, Dunn (1968) observed that minority groups were over-identified as handicapped. He believed this overrepresentation was caused by discriminatory intelligence and achievement tests. Dunn and many others (e.g., Cole & Bruner, 1971; Gould, 1981; McClelland, 1973; Ogbu, 1978) have contended that these tests are biased primarily because they are ethnocentric: Test content is drawn exclusively from white middle class experience. Evidence for minority disproportionality in special education and claims of biased testing were influential in several heralded court cases in the 1970s (e.g., Larry P. v. Wilson Riles, 1971), which curtailed intelligence testing in many school districts (see Bersoff, 1981).

Nevertheless, psychometricians are increasingly skeptical that many well-known and widely used intelligence and achievement tests are biased against minorities. Reschly (1981), for example, has pointed out that subjective judgment, rather than data, often has been the basis for charges that these tests are ethnocentric. More importantly, recent empirical investigations of such tests' content, construct, and criterion validity have failed to show that they are biased (e.g., Cole, 1981; Gutkin & Reynolds, 1980; Jensen, 1980; Linn, 1982; Oakland, 1983; Reynolds, 1982; Sandoval, Zimmerman, & Woo-Sam, 1980), leading Reynolds (1983) to conclude that, "Empirical research into the question of bias has failed to substantiate the existence of cultural bias in well constructed, well standardized educational and psychological tests when used with native-born American ethnic minorities."

There is something remarkable about this intense, sustained, and well publicized debate: Both sides tend to focus narrowly on the test instrument and virtually ignore the context in which assessment occurs. Research has been infrequent to nonexistent with respect to contextual factors such as (a) examinees' interpretation of the purpose of testing and comprehension of test instructions and (b) examiners' personality, reinforcement strategies, pretest information on examinees, attitudes about the legitimacy of testing, order in which tests of varying difficulty are administered, and choice of test location. This paucity of research on context is not surprising, given that we tend to conceptualize the test situation as decontextualized, as a setting in which extra-test factors can be controlled and their effects on performance neutralized (cf. Mehan, 1978; Sigel, 1974). Surprising or not, our uninterest in test context prevents us from knowing whether typical situational factors in testing affect minority and non-minority children differently.

One exception to the foregoing is the specific question of whether black children achieve higher scores when tested by black, rather than by white, examiners, an issue receiving moderate attention by researchers (cf. Sattler & Gwynne, 1982). Another contextual variable explored with relative frequency, although not with respect to minority assessment, is examiner familiarity.

Fuchs and associates have demonstrated that handicapped children obtain higher scores when tested by familiar, rather than by unfamiliar, examiners and that this pattern of performance appears robust. Differential performance in favor of the familiar examiner was obtained (a) when examiners were inexperienced and when they were highly trained and experienced professionals, (b) across studies

employing experimentally induced and long-term acquaintanceship definitions of familiarity, (c) over various levels of item difficulty and response modes, (d) irrespective of the sex of examinees, (e) for both preschool and school-age language-handicapped students, (f) among language-impaired and learning disabled populations, and (g) regardless of whether examinees' performance was scored by the examiner or a "blind" rater responding to a videotaped replay of the testing (Fuchs, Featherstone, Garwick, & Fuchs, 1984; Fuchs & Fuchs, 1984; Fuchs, Fuchs, Dailey, & Power, 1985; Fuchs, Fuchs, Garwick, & Featherstone, 1983; Fuchs, Fuchs, Power, & Dailey, 1983; Fuchs, Fuchs, Power, Duval, & Sacco, 1986)). A recent study has shown that unfamiliar examiners depress the performance of handicapped, but not nonhandicapped, children (Fuchs, Fuchs, Power, & Dailey, 1985), indicating that examiner unfamiliarity is a source of systematic error or bias in the assessment of handicapped children. The importance of this finding is underscored by the fact that most examiners in schools and clinics are strangers to their examinees (Fuchs, 1981).

Because examiner unfamiliarity is part of the test procedure, rather than the test instrument per se, we choose to refer to this systematic error as "test procedure bias." Given that an unfamiliar examiner appears to negatively bias the test procedure with handicapped children, one may ask whether examiner unfamiliarity constitutes a similar bias against minority pupils. If so, then the ubiquitous procedure of employing unfamiliar examiners contributes to a spuriously low performance of minority children and increases the likelihood that they will be identified inaccurately as handicapped. Such a possibility should be of concern to school psychologists and administrators, test developers and publishers, and those who set

professional standards for testing as well as parents and teachers of minority students. Thus, a wide-ranging quantitative synthesis was conducted of the examiner unfamiliarity literature to determine the importance of this contextual factor to minority (i.e., black and Hispanic) and Caucasion students.

Method

Search Procedure

The search for pertinent studies was conducted primarily by a computer search of three on-line data bases: ERIC (from 1966), Psych Info (from 1967), and Dissertation Abstracts (from 1927). Additionally, a manual search was conducted of 12 educational, psychological, and speech/language journals (1965-1982, inclusive) and the reference sections of selected textbooks. Finally, titles in the references of all identified investigations were pursued.

A study was considered for inclusion if it compared examiner familiarity to unfamilarity in terms of effects on examinees' performance during individualized testing. The search yielded 22 studies, of which 14 provided unambiguous data on Caucasion and/or minority examinees' performance in familiar and unfamiliar examiner conditions. Of these 14 studies, 6 involved only Caucasion children, 6 included only minority (black and/or Hispanic) children, and 2 employed groups of both Caucasion and minority subjects. Thus, an equal number of studies (N =8) provided data on minority and Caucasion pupils' performance in the two examiner conditions.

Data Extracted from Each Study

Results of the 14 studies were transformed to a common metric, effect size. Effect sizes were derived by determining the mean difference between examinees' scores in the familiar and unfamiliar

examiner conditions and dividing this difference by the standard deviation of examinees' scores in the unfamiliar condition (Glass, McGaw, & Smith, 1981). For studies reporting relevant means and standard deviations for examinees' performance in familiar and unfamiliar examiner conditions, effect sizes were calculated from these statistics; for studies not reporting means and standard deviations, effect sizes were calculated from other statistics such as F or p values (see Glass et al., 1981). Some of the 14 studies reported more than one effect. In all but two instances, a median effect size of examiner familiarity/unfamiliarity was calculated for each study. The exceptions were the two investigations incorporating separate groups of Caucasion and minority examinees within the same experimental design. In each of these studies two effect sizes were reported, one for minority examinees and one for Caucasion examinees. Thus, a total of 16 effect sizes (8 for Caucasion children an 8 for minority children) was derived from the 14 studies. Each effect size was converted to unbiased effect sizes (UESs), correcting for the inconsistency in estimating true from observed effect sizes (Hedges, 1981). In combining these UESs, weighted averages were calculated to account for the variance of the UESs (see Hedges, 1984).

Methodological Study Features

Effects of examiner familiarity/unfamiliarity were related to one composite procedural variable. The composite procedural variable indicates the overall methodological quality of each investigation. It was based on an analysis of nine design-related features. These methodological features, as well as the standards against which they were judged to generate an overall quality index, follow.

1. Assignment of examinees to examiners. It was necessary for

examinees to be assigned randomly to examiners.

2. Assignment of examinees to treatments. Investigators were required to assign examinees randomly to experimental conditions, or to use a repeated measures design.

3. Examiner expectancy. Researchers were expected to insure that examiners were blind to the general experimental questions and, specifically, to the familiar/unfamiliar nature of the test conditions.

4. Fidelity of treatment conditions. Investigators employing a personal acquaintanceship definition of familiarity were required to make explicit that unfamiliar examiners were strangers to examinees and that examiner familiarity either represented a long-term acquaintanceship between test participants or was the resultant of an experimentally-induced procedure.

5. Multiple treatment effects. Studies were evaluated as acceptable when effects of the familiar/unfamiliar conditions did not appear to be confounded with other factors such as the gender of familiar and unfamiliar testers.

6. Number of examiners. It was judged important that there be a minimum of two familiar and two unfamiliar examiners.

7. Order of testing. Studies employing a repeated measures design were required to counterbalance testing in familiar and unfamiliar examiner conditions.

8. Scoring. It was necessary that scores be calculated by a blind procedure.

9. Technical adequacy of dependent measure. At a minimum, a study was expected to use measures with indices for internal or test-retest reliability exceeding .69.

Interrater agreement[1] on each of these dimensions, based on two raters' scores on six randomly selected studies (43% of the sample), ranged from .67 to 1.00. Average agreement across all nine methodological characteristics was .89.

Methodological Quality of Studies

Since 1 of the 14 studies provided insufficient information to determine methodological quality, the quality of 13 studies was quantified employing a four-step procedure. First, every investigation was analyzed in terms of the nine design-related features and criteria described above. These design features were coded acceptable (1 point), unacceptable (0 points), or not applicable. Second, a weight of 1 or 2 was assigned to each methodological charactreristic. "Technical adequacy of dependent measure," "assignment of examinees to treatments," and "assignment of examinees to examiners" received a weight of 2; the remaining six design characteristics received a weight of 1. Third, a composite score was generated for each study by multiplying the coded values (1 or 0) by the assigned weights (1 or 2), summing these products, and then dividing the sum by the number of applicable study characteristics. Finally, a frequency distribution of these composite scores was generated. It indicated that 7 investigations received a composite score between 1.00 and 1.43 (high quality); 6 studies were assigned composite scores between .33 and .80 (low quality).

Results

A test for the homogeneity of effect size (Hedges, 1982), undertaken to determine whether the population effect size was constant across Caucasion and minority unbiased UESs, yielded a significant value, $\chi(15, N = 16) = 89.22$, $p < .01$. Therefore,

additional analyses were conducted to explain variations in UESs by examinees' Caucasion/minority status.

To compare magnitude of UESs of Caucasion and minority examinees, Hedges's (1984) chi square analogue to analysis of variance was employed. With conventional analysis of variance conducted on effect sizes, problems exist because it is possible for systematic variance to be pooled into the estimate of error variance. Moreover, violation of the homoscedasticity assumption is severe in research synthesis, and there is little reason to believe that the usual robustness of the F test will prevail (see Hedges, 1984). The use of Hedges's analogue to analysis of variance avoids these conceptual and statistical problems.

Methodological Quality of Studies with Caucasion and Minority Examinees

The mean quality rating for the 8 studies involving Caucasion examinees was .99 (SD = .40); the average quality rating for 7 studies associated with minority examinees was .91 (SD = .40). This difference was not statistically significant, t (13) = .39, ns.

Comparing Caucasion and Minority Examinees' Performance

For Caucasion examinees, the average weighted UES was .05 (v = .073), z = .72, ns. The average weighted UES for black and Hispanic examinees was .72 (v = .096), z = 7.47, $p < .001$. A chi square analogue to analysis of variance indicated that this difference was statistically significant, $\chi(1, N = 16) = 30.35$, $p < .001$. The minority group's UES indicates that, given a normative test (such as an intelligence measure) with a population mean of 100 and a standard deviation of 15, the use of a familiar examiner would raise the typical minority student's score from 100 to 111. In

contradistinction, the Caucasion group's UES suggests virtually no change in score as a function of examiners' familiarity/unfamiliarity. In terms of Cohen's (1977) well known U_3 (or percentage of nonoverlap) statistic, the upper 50% of the minority students' distribution of scores in the familiar examiner condition exceeded 76% of the distribution of scores in the unfamiliar examiner condition (see Figure 1). For Caucasion examinees, their distribution of scores in the familiar condition nearly was superimposed on the distribution of scores in the unfamiliar condition.

Insert Figure 1 about here

Discussion

Whereas Caucasion students performed similarly in familiar examiner and unfamiliar examiner conditions, black and Hispanic children scored significantly and dramatically higher with familiar examiners. This indicates examiner unfamiliarity selectively depresses the performance of black and Hispanic examinees and represents test procedure bias in the assessment of minority children. This conclusion, of course, must be tempered by the fact that it is based on a quantitative synthesis of 14 empirical studies and one legitimately may question the stability and generalizability of the data base. However, if we assume that the data are representative, then they have several important implications for practice and research.

Practically, test developers' use of unfamiliar examiners to generate normative data and indices of validity (cf. Fuchs, Fuchs, Dailey, & Power, 1983) appears problematic for minority pupils.

Comparing minority students' presumably suboptimal performance with unfamiliar examiners to the more maximal performance of largely Caucasion normative populations could result in spuriously low and improperly restrictive educational placements of minority children. Indeed, examiner unfamiliarity may be a partial explanation for the frequently noted overrepresentation of minorities in special education classrooms. If this is so, then examiner unfamiliarity represents a condition under which disproportionality of placement constitutes inequity of treatment, as defined by the National Research Council's Panel on Selection and Placement of Students in Programs for the Mentally Retarded (cf. Messick, 1984). The apparent bias caused by the use of unfamiliar examiners also is an explicit violation of Section 615-5c of PL 94-142, which states, "testing and evaluation materials and procedures utilized for the purposes of evaluation and placement of handicapped children will be selected and administered so as not to be racially or culturally discriminatory."

The foregoing underscores the view that testers should be familiar with minority children prior to testing. Even if testers might view such a prescription as conceptually sound, many also might consider it infeasible given the severe time constraints under which they frequently operate (see AERA, APA, & NCME, 1983, p. 14-2). Yet if it were possible to predict which minority students are more likely to perform suboptimally with an unfamiliar examiner, then testers might establish pretest contact with only a subgroup of pupils. Recently Fuchs, Fuchs, and Blaisdell (1986) attempted such a prediction for language-handicapped students on the basis of information gathered from multiple sources (e.g., teachers, peers, and subjects) and by qualitatively different methods (e.g., teacher

ratings, peer nomination, and self-report). Four predictor variables accounted for 39% of the variance in the differential performance of the handicapped examinees. Whether these predictors are similarly efficient for minority students is an empirical issue that future research might address.

In a similar research vein, examiner unfamiliarity may partly explain why, on average, minority children obtain lower IQ scores than Caucasion children. A frequent estimate of the magnitude of this difference in IQ performance has been one standard deviation (cf. Linn, 1982). Minority children's test performance conventionally has been interpreted as a rather straightforward demonstration of those skills and abilities that the tests claim to measure. Typically, the minority students' comparatively poor showing on these tests has been attributed primarily to poor genes or a disadvantaged environment (see Nichols, 1978). Nevertheless, current findings question such interpretations that presume a cause and effect relation between children's cognitive processes and their performance on tests that purportedly measure salient cognitive and/or academic abilitites. Our results indicate that at least one extra-test factor, examiner unfamiliarity, also affects the performance of select groups of children. For minority pupils, the effect size associated with examiner familiarity was .72, which is the equivalent of a difference of approximately 11 points on a standardized IQ test with a mean of 100 and a standard deviation of 15.

Examiner familiarity's sizable influence on black and Hispanic examinees' performance also has implications for our understanding of another contextual variable, perhaps the most frequently investigated extra-test factor in minority assessment; namely, race of examiner

effects. Sattler and Gwynne (1982) summarized 27 studies of this issue and concluded that, contrary to popular belief, there is little empirical evidence that white examiners adversely affect the test performance of black examinees. However, Graziano, Varca, and Levy (1982) reviewed much of the same literature and reached a somewhat different conclusion. Graziano et al. observed that, taken as a whole, the pertinent studies neither (a) provide strong evidence that examiners of different races systematically elicit different performance in black and white examinees, nor (b) lay to rest the issue of examiner's race. Graziano et al. claimed that one source of this confusion is the traditionally narrow conceptualization of the race of examiner problem: Examiner's race has been treated as a "macrovariable," with all white examiners seen as interchangeable and all black examiners as interchangeable. Future research, argued Graziano et al., must be more analytical.

Findings from this quantitative synthesis are consonant with the Graziano et al. argument. Examiner unfamiliarity's depressing effect on minority examinees' performance, together with the probable fact that unfamiliar examiners were used in most studies of race of examiner effects, suggests that the influence of examiners' race frequently may have been masked by the experimental procedure (i.e., unfamiliar examiners) employed. A plausible, although untested, hypothesis is that examiners' race becomes salient only after examinees and examiners become personally acquainted.

Finally, results of this investigation indicate that it is precipitous, if not incorrect, to claim testing is unbiased toward minority children. Whereas many psychological and educational test instruments may not be biased, at least one facet of the test

procedure, examiner unfamiliarity, appears to depress selectively minority examinees' performance. Furthermore, there may be additional contextual variables constituting the typical test situation that bias against certain pupils' performance. This suggestion might strike some psychometricians as unorthodox, but it is by no means new. Fifteen years ago, Cronbach (1971) stated, "The (test) instrument...is only one element in a procedure, and a validation study examines the procedure as a whole" (p.449).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1983, February). Draft: Joint technical standards for educational and psychological testing. (Available from American Psychological Association, Office for Scientific Affairs, 1200 17th Street, N.W., Washington, DC 20036)

Bersoff, D.N. (1981). Testing and the law. American Psychologist, 36, 1047-1056.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cole, N.S. (1981). Bias in testing. American Psychologist, 36, 1067-1077.

Cole, M., & Bruner, J.S. (1972). Preliminaries to a theory of cultural differences. In I. Gordon (Ed.), Seventy-first yearbook of the National Society for the Study of Education, part II--Early childhood education. Chicago: University of Chicago.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), Educational measurement (pp.443-507). Washington, DC: American Council on Education.

Fuchs, D. (1981, April). Differential responses of preschool language-handicapped children to familiar and unfamiliar testers as a function of task complexity, length of acquaintanceship, and sex of child. In V. Shipman (Chair), Client identification and issues of validity: The influence of situational variables on children's cognitive performance. Symposium presented at the annual meeting of the American Educational Research Association, Los Angeles.

Fuchs, D., Featherstone, N., Garwick, D.R., & Fuchs, L.S. (1984). Effects of examiner familiarity and task characteristics on speech and

language-impaired children's test performance. Measurement and Evaluation in Guidance, 16, 198-204.

Fuchs, D. Fuchs, L.S., & Blaisdell, M. (1986). Psychosocial characteristics of handicapped children who perform suboptimally during assessment. Measurement and Evaluation in Counseling and Development, 18, 176-184.

Fuchs, D., Fuchs, L.S., Dailey, A.M., & Power, M.H. (1983). Effects of pretest contact with experienced and inexperienced examiners on handicapped children's performance (Research Report No. 110). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Fuchs, D., Fuchs, L.S., Dailey, A.M., & Power, M.H. (1985). The effect of examiners' personal familiarity and professional experience on handicapped children's test performance. Journal of Educational Research, 78, 141-146.

Fuchs, D., Fuchs, L.S., Garwick, D.R., & Featherstone, N. (1983). Test performance of language-handicapped children with familiar and unfamiliar examiners. Journal of Psychology, 114, 37-46.

Fuchs, D., Fuchs, L.S., Power, M.H., & Dailey, A.M. (1983). Systematic bias in the assessment of handicapped children (Research Report No. 134). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities. (ERIC Document Reproduction Service No. 236 200)

Fuchs, D., Fuchs, L.S., Power, M.H., & Dailey, A.M. (1985). Bias in the assessment of handicapped children. American Educational Research Journal, 22, 185-198.

Fuchs, D., Fuchs, L.S., Power, M.H., Duval, N., & Sacco, L. (1986). Importance of context in testing children of different cognitive competence. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Fuchs, L.S., & Fuchs, D. (1984). Examiner accuracy during protocol completion. Journal of Psychoeducational Assessment, 2, 101-108.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Gould, S.J. (1981). The mismeasure of man. New York: W.W. Norton.

Graziano, W.G., Varca, P.E., & Levy, J.C. (1982). Race of examiner effects and the validity of intelligence tests. Review of Educational Research, 52, 469-497.

Gutkin, T.B., & Reynolds, C.R. (1980). Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services. Journal of School Psychology, 18, 34-39.

Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 359-361.

Hedges, L. (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92, 490-499.

Hedges, L. (1984). Advances in statistical methods for meta-analysis. In W.H. Yeaton & P.N. Wortman (Eds.), Issues in data synthesis. New Directions for Program Evaluation, 24, San Francisco: Jossey-Bass.

Jensen, A.R. (1980). Bias in mental testing. New York: The Free Press.

Larry P. v. Wilson Riles, (1971). U.S. District Court for the Northern District of California, No. C-71-2270 RFP.

Linn, R. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A.K. Wigdor & W.R. Garner (Eds.), Ability testing: Uses, consequences, and controversies. Washington, DC: National Academy Press.

Mehan, H. (1978). Structuring school structure. Harvard Educational Review, 48, 32-64.

Messick, S. (1984). Assessment in context: Appraising student performance in

relation to instructional quality. Educational Researcher, 13, 3-8.

McClelland, D.C. (1973). Testing for competence rather than for "intelligence." American Psychologist, 28, 1-14.

Nichols, R.C. (1978). Policy implications of the IQ controversy. Review of Research in Education, 6, 3-46.

Oakland, T. (1983). Concurrent and predictive validity estimates for the WISC-R IQs and ELPs by racial-ethnic and SES groups. School Psychology Review, 12, 57-61.

Ogbu, J.U. (1978). Minority education and caste. New York: Academic Press.

Reschly, D.J. (1981). Psychological testing in educational classification and placement. American Psychologist, 36, 1094-1102.

Reynolds, C.R. (1982). The problem of bias in psychological assessment. In C.R. Reynolds & T.B. Gutkin (Eds.), The handbook of school psychology, New York: Wiley.

Reynolds, C.R. (1983). Test bias: In God we trust; all others must have data. The Journal of Special Education, 17, 241-260.

Sandoval, J., Zimmerman, J.L., & Woo-Sam, J. (1980). Cultural differences on WISC-R verbal items. Paper presented at the annual meeting of the American Psychological Association, Montreal.

Sattler, J.M., & Gwynne, J. (1982). White examiners do not impede the intelligence test performance of black children: To debunk a myth. Journal of Consulting and Clinical Psychology, 50, 196-208.

Sigel, I.E. (1974). When do we know what a child knows? Human Development, 17, 201-217.

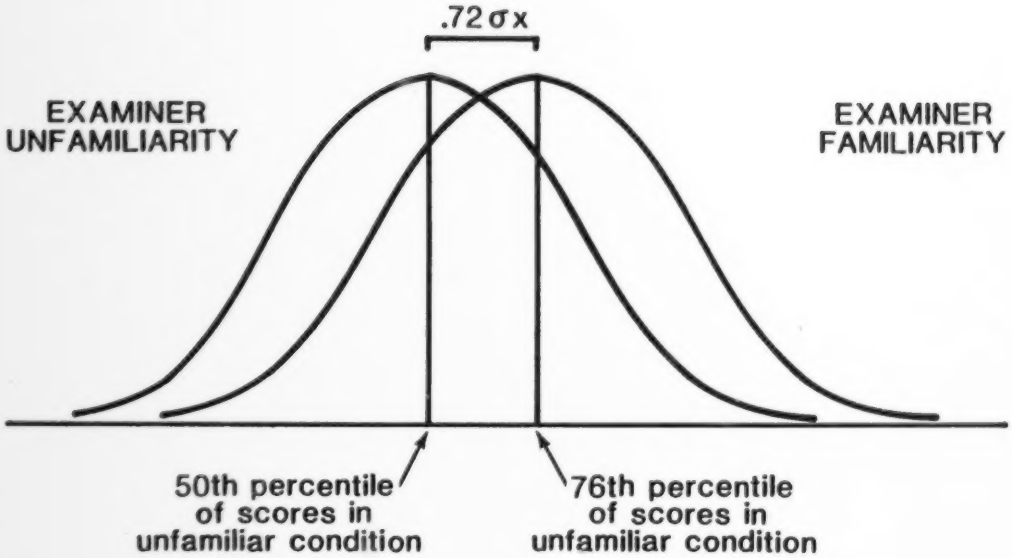
Thompson, R.H., White, K.R., & Morgan, D.P. (1982). Teacher-student interaction patterns in classrooms with mainstreamed mildly handicapped students. American Educational Research Journal, 19, 220-236.

Footnote

[1] Interrater agreement was calculated using the following formula (Coulter cited in Thompson, White, & Morgan, 1982): Percentage of agreement = agreements between raters A and B / (agreements + disagreements between raters A and B + omissions by rater A + omissions by rater B.

Figure Caption

Figure 1. Effect of examiner familiarity on black and Hispanic students' test performance.


.72σx
EXAMINER
UNFAMILIARITY
EXAMINER
FAMILIARITY
50th percentile
of scores in
unfamiliar condition
76th percentile
of scores in
unfamiliar condition